**Oncology Data Representation Challenges**

**Rimma Belenkaya, Michael Gurley, and OHDSI Oncology Workgroup**

**November 18, 2017**

1. **Key research and clinical questions**

* Identification and follow-up of patients with a certain disease phenotype (histology, location, grade, staging, etc.)
* Identification of treatment regimen and response to treatment
* Identification of recurrences and progression of disease
* Measuring and comparing survival time from diagnosis
* Prediction of recurrences, length of remissions, end of life events

1. **Data standardization challenges**

We have identified three major types of challenges in standardizing oncology data. First is identifying and extracting critical data elements from the source data. Second is representing these data structurally and terminologically in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) maintained by Observational Health Data Sciences and Informatics (OHDSI). Third is analytical derivation of the key disease features that cannot be extracted directly from the source data.

**Source data challenges**

1. Identifying cancer diagnosis and recurrences
   1. Precision of cancer diagnosis

ICD-O is a gold standard to annotate cancer diagnosis. It is only available for the first cancer occurrence in most (non-SEER state) cancer registries. It is not available in electronic medical records (EMRs). In EMRs, cancer diagnosis is recorded using ICD-9/10 coding that is less granular than ICD-O.

* 1. Tracking cancer patients throughout the course of disease

Cancer occurrences are not available in a structured form in EMRs. ICD -9/10 diagnoses in EMRs may change throughout the course of the disease. Connecting ICD-O diagnosis in cancer registry with ICD-9/10 diagnoses in EMRs and tracking the same cancer throughout the course of the disease presents a significant challenge. Patterns of ICD9/10 diagnoses after initial diagnosis do not reliably track the recurrence status of a patient’s cancer diagnosis.

1. Staging, grade, and other cancer specific attributes.
   1. Pathology staging in a structured form is available only in cancer registry for the first cancer occurrence. Clinical staging is also available in cancer registry, but is of known poor quality. AJCC staging versions have overlapping ambiguous coding.
   2. Other cancer attributes are represented as structured data in cancer registries for the first cancer occurrence only (except for SEER states).
   3. These data are rarely defined as structured coded elements in EMRs.
2. Identifying treatment and treatment regimens

Cancer treatments are usually delivered as a regimen combining one or more treatment modalities (e.g. chemotherapy, radiotherapy, surgery) and administered through treatment cycles. Identifying the whole regimen, its components, their order and periodicity is critical.

* 1. Treatment regimens are represented only in cancer registries, usually without specific treatment details (except for SEER states). For example, chemotherapy followed by surgery. Treatment details (e.g. medications, surgeries) are usually stored in EMRs.
  2. Most EMR systems do not indicate treatment intent. Therefore, separating cancer treatments from other treatments is a substantial challenge, especially in general hospitals. Radiotherapy is the least ambiguous treatment, systemic treatments and surgeries are more difficult to identify.
  3. Some EMR systems have medications regimens recorded in order sets, but it is not always the case. Treatment cycles are not normally recorded in a structured way.

1. Identifying response to treatment and disease progression

These key cancer observations are not usually recorded in a structured way in any source including EMRs and cancer registries.

* 1. Imaging response to treatment (RESIST) is available in clinical trials for solid tumors.
  2. Change of treatment is usually used as a proxy for detecting disease progression. However, it may also indicate a planned change or change due to toxicity.

**Modeling and terminology challenges**

Although OMOP CDM requires some structural extensions to support oncology data, these will be easily accomplished using available community resources. The primary challenge is in choosing, validating, and extending most suitable terminologies to represent oncology domain.

1. Diagnosis representation

As described above, cancer diagnosis is represented in ICD-O for the first occurrence and in ICD-9/10 for all occurrences. The challenge is to connect these two representations. OHDSI’s standard for diagnosis representation is SNOMED. ICD-9 and ICD-10 have been successfully mapped to SNOMED. The current challenge is to validate existing mappings between ICD-O and SNOMED CT and propose new SNOMED CT coding to cover all the existing ICD-O histology and topography combinations.

1. Staging, grade, and other cancer specific attributes.

Although many of these characteristics are covered in LOINC and SNOMED CT, the challenge is to create a set of these characteristics for each group of cancer diagnosis (e.g. breast, lung, etc.). There is an ongoing initiative to create and align these sets with the ICCR (International Collaboration on Cancer Reporting) data sets. However, at this point, it has only covered a few cancer types.

1. Treatment representation

Overall, OMOP CDM will have to be extended to represent cancer treatment regimens, including treatment modalities (systemic, surgery, radiotherapy, palliative care) and cycles. Unlike in other disease domain, in oncology this information may be either derived or extracted directly from the source data. This challenges OMOP delineation between derived and directly extracted data which will have to be resolved.

* 1. Systemic treatments (chemo-, immuno-, hormone, and targeted therapies)

First challenge in this domain is to provide drug combinations that represent standard of care regimens so that these regimens can be identified in the source data and then represented as such in OMOP CDM. Such lists are available through SEER, without linkages to a specific cancer type, and through NCCN, including linkages to a specific cancer type. The latter is a licensed content that cannot be easily obtained by OHDSI.

Second challenge is representing these drug regimens in OMOP CDM. Current drug era tables do not provide the required structure, but may be supported by the fact\_relationship table.

* 1. Surgery

The challenge in this treatment domain is identifying cancer-related surgeries. CPT codes that normally represent surgeries do not have classifications that would help identifying them as cancer specific. OMOP vocabulary provides linkages between CPT and SNOMED CT. However, the quality of these mappings has to be validated.

* 1. Radiotherapy

Radiotherapy treatment representation also needs classification (e.g. gamma, stereotactic, etc.) that is not available for CPT/HCPCS codes. The challenge is similar to the surgery challenge: CPT/HCPCS to SNOMED CT mappings as well as SNOMED CT classifications must be validated.

Additionally, to have a complete representation of radiotherapy (e.g. dosage, frequency, etc.) and surgery (e.g. laterality) requires an extension of the current OMOP CDM procedure domain.

* 1. Palliative care

Vocabulary for this domain will have to be identified.

1. Genomic data currently does not have either vocabulary or model support in OMOP CDM. They have to be added to the model. In order to accomplish this, we need domain experts (geneticist, oncologist) to join the effort.

**Analytical derivation of the key disease features challenge**

As indicated above, a lot of key disease features including treatment regimens, response to treatment, disease progression and recurrences will not be available in the source data. All these will have to be derived. Additional treatment classifications and groupings in the vocabulary may support these algorithms. Data available from the sources will help validating these algorithms.

1. **Specific needs for centralization and convergence of Oncology treatment compendium efforts.**

All distribution of compendiums should be simple and stable; CSV, XML, and/or JSON serialization and stable identifiers should be available for all items.

**Drugs**

* One list of Oncology drugs, normalized to RxNorm, classified at an intuitive level of abstraction(Chemotherapy, Hormonal Therapy, Immunotherapy), connected to coded known/approved evidentiary indications ICD-10/ICD-O and regimen participation.
* One list of Oncology drugs procedure codes connected to a procedure vocabulary standard (HPCS), classified at an intuitive level of abstraction (Chemotherapy, Hormonal Therapy, Immunotherapy), connected to coded known/approved evidentiary indications ICD-10/ICD-O and regimen participation.
* One list of Oncology drug regimens, normalized to an appropriate standard, delineating drug participation.

<https://seer.cancer.gov/seertools/seerrx/>

<https://www.cancer.gov/about-cancer/treatment/drugs>

<https://crn.cancer.gov/resources/codes.html>

<https://ncim.nci.nih.gov/ncimbrowser/>

<https://www.nccn.org/professionals/drug_compendium/content/>

**Radiotherapy**

* One list of radiotherapy procedure codes (HPCS/CPT), classified at an intuitive level of abstraction in a standardized procedure vocabulary.
* Normalization of radiotherapy treatment modifiers to a standardized vocabulary, radiotherapy modality, fractions, dose, anatomical targets of treatment.

<https://crn.cancer.gov/resources/codes.html>

<https://ncim.nci.nih.gov/ncimbrowser/>

<https://www.nccn.org/professionals/radiation/content/>

**Surgeries**

* One list of surgical procedure codes (HPCS/CPT), classified at an intuitive level of abstraction (Biopsy, Resection) in a standardized procedure vocabulary.

**Tumor Registry**

* Mapping of all NAACCR/SEER variables to standardized vocabularies.

<https://www.naaccr.org/data-standards-data-dictionary/>

**Tumor Characteristics/Genomics**

* Help accelerate Prof. William Campbell’s efforts to encode all the CAP Cancer Checklists anatomic and molecular pathology findings/observables to a standardized vocabulary.

<https://www.unmc.edu/pathology/informatics/tdc>

<https://www.unmc.edu/pathology/faculty/bios/campbell.html>

**Recurrence/Progression/Response**

* Support creation of standardized site-specific phenotype algorithms capable of detecting/deriving cancer recurrence/treatment response/disease progression from discrete clinical data and clinical narratives.